PB82-264961

Statistical Basis for Laboratory Performance Evaluation Limits

(U.S.) Environmental Monitoring and Support Lab.
Research Triangle Park, NC

Aug 82

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3.
EPA-600/D-82-300	PUBLISHED PAPER	PB82-264961

4. TITLE AND SUBTITLE	5. REPORT DATE
Statistical Basis for Laboratory Performance Evaluation Limits	August 1982
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Paul W. Britton	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
Quality Assurance Branch EMSL-Cincinnati, U.S. Environmental Protection Agency 26 W. St. Clair Street Cincinnati, OH 45268	ABLC1A, AAPB1A, CBEB1C
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
Quality Assurance Branch EMSL-Cincinnati, U.S. Environmental Protection Agency 26 W. St. Clair, Street Cincinnati, OH 45268	Proceedings
	14. SPONSORING AGENCY CODE EPA 600/6

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The U.S. Environmental Protection Agency (EPA) conducts studies to evaluate the performance of drinking water and wastewater laboratories that analyze samples for major EPA programs. The studies involve sample concentrates which the participating laboratories dilute to volume with reagent water. Analytical results are evaluated by comparison with confidence limits.

From completed studies the characteristics of data representing proper laboratory operation are estimated using robust procedures on the data from EPA and state laboratories. From such historical statistics and their related true sample concentrations, weighted linear regressions are developed and used to predict the statistics for subsequent studies. If these regression predictions are similar to the statistics actually calculated from the current study data, the evaluation limits are calculated from the regression statistics for the sake of independence and to force a smooth relationship between the limits at different sample concentration.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
Distribute to Public	Unclassified	9
	20. SECURITY CLASS (This page)	22. PRICE
	Unclassified	

EPA-600/D-82-300
August 1982

STATISTICAL BASIS FOR LABORATORY PERFORMANCE EVALUATION LIMITS

Paul W. Britton
Quality Assurance Branch
Environmental Monitoring and Support Laboratory - Cincinnati

1. INTRODUCTION

The effectiveness of the water regulations enforced by the U.S. Environmental Protection Agency (USEPA) depends heavily upon the quality of data generated by supporting laboratories. Most of the data are generated by state, local-government, industrial, commerical, or other non-USEPA laboratories. In an effort to improve the quality of data, USEPA has begun mandatory collaborative studies evaluating the ability of laboratories to analyze water samples and to produce data within specific acceptance limits. Acceptable performance demonstrates a well-managed, competent laboratory operating at its full potential, while unacceptable performance results in an appropriate remedial response by USEPA.

During these performance evaluation (PE) studies, the Quality Assurance Branch of the Environmental Monitoring and Support Laboratory - Cincinnati sends participating laboratories a set of stable sample concentrates in sealed glass ampuls, a data reporting form and appropriate instructions. Each laboratory produces the study samples by diluting a measured quantity of specific concentrates to volume with reagent water, then analyzes them using its routine procedure. The completed form is sent to USEPA for evaluation and a fully-detailed report is returned to each laboratory. The responsible state of USEPA office follows up with laboratories that demonstrate problems.

The effectiveness of these studies depends upon two things: first, whether the acceptance limits properly identify data resulting from substandard analytical work; and second, the commitment of all participants to ensure that laboratories submitting unacceptable results either improve their performance or are excluded from producing further data.

2. DEVELOPMENT OF ACCEPTANCE LIMITS

EPA's objective in developing acceptance limits is to be able to distinguish between data from analytical systems operating within state-of-the-art capabilities and data representing substandard performance. A statistical confidence interval at an appropriate significance level is clearly suitable if acceptable data have a normal distribution and if proper estimates of the arithmetic mean and standard deviation are available. Since there are no independent sources of such information, it is necessary to investigate the data from the studies themselves to establish a basis for the desired limits.

2.1 Estimating the Study Statistics

For each chemical constituent in a sample, basic descriptive statistics are estimated from the data submitted by the participating USEPA and state laboratories, which usually number over 100. The statistics are estimated from these data because USEPA is more familiar with the laboratories and is confident that they will produce results generally representative of the current analytical procedures when properly applied. The report for each chemical constituent includes the following:

1. Identification of obvious outliers in the data set;

2. Traditional and robust estimates of the mean and standard deviation from the retained data;

3. Tests of normality for the retained data;

4. A histogram of the retained data; and

5. A complete listing of the USEPA and state data with the obvious outliers noted.

Outliers are identified by a two-stage procedure. The first stage involves the rejection of values larger than 500 percent or less than 20 percent of the true value and is intended to remove all data resulting from a decimal placement error. Data retained through this first stage are used to calculate traditional preliminary estimates of the mean and standard deviation for the second stage tests. If there are 25 or fewer values left, the second stage involves successive Grubbs' tests at the two-tailed 1 percent significance level until a suspected value is retained. As described in Grubbs (1950 and 1969) and Standard Practice E178 of the American Society for Testing and Materials, this test involves ordering the data, identifying the largest or smallest value as the suspected value X --- the value farthest from the mean of all the data --- and calculating $T = (X - \overline{X})/S$, where $\overline{X}$ is the estimate of the mean and S is the estimate of the standard deviation. The suspected value is discarded as an outlier if the absolute value of T exceeds the appropriate critical value; then $\overline{X}$ and S are recalculated from the remaining data, prior to testing the next most extreme value. For sets of more than 25 retained values, any value beyond a 99 percent confidence interval, generated from the appropriate Student's t value, is considered an outlier.

Final estimates of the mean and standard deviation are calculated from the n values retained through both stages of outlier testing. The traditional estimates are:

$$\bar{X} = \left(\sum_{i=1}^{n} X_i \right)/n \quad , \tag{1}$$

$$S = \sqrt{\frac{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2}{n(n-1)}} \quad , \tag{2}$$

where X_i is the ith retained value.

Robust estimates are calculated from an ordered set of the retained data, using a 25 percent Winsorizing procedure to estimate the mean as described by Dixon and Massey (1969), and standard normal deviates to estimate the standard deviation.

After the retained data have been arranged in ascending order, the Winsorized estimate of the mean is calculated as:

$$\bar{X} = \left[K(X_K) + \sum_{i=K+1}^{n-K} X_i + K(X_{n-K+1}) \right] /n \quad , \tag{3}$$

where K = (the greatest integer portion of n/4) + 1.

The robust estimate of the standard deviation is calculated from the interquartile distance across the middle 50 percent of the ordered retained data. Assuming an underlying normal distribution, this distance is equivalent to 1.349 standard deviations. The calculation is:

$$S = \left(\frac{1}{1.349}\right) \left[X_{n-J+1} - X_J + \left(J - \frac{n}{4} - .5\right)\left((X_J - X_{J-1}) + (X_{n-J+2} - X_{n-J+1})\right) \right], \tag{4}$$

where J = (the greatest integer portion of (n+1)/4) + 1.

To test normality, Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) statistics are calculated for the full set of retained data and a K-S statistic is calculated for the middle 50 percent of the retained data. The K-S statistic is used because it is a highly regarded distance statistic that can be used to test normality of the middle 50 percent of the retained data as well as the full set. The A-D statistic is calculated for the full set of retained data because it is more sensitive to deviations from normality in the distribution tails, and was found to be one of the most powerful and easily calculated distribution tests studied by Stevens (1974) and Green and Hegazy (1976). Suitable critical values are provided at 1 and 5 percent significance levels using the approximation calculations provided by Stevens (1974).

For parameters with rugged analytical procedures or with which the laboratories have considerable experience, there is no practical difference between the traditional and robust estimates of the mean and standard deviation. But those parameters for which a number of the USEPA and state laboratories still do not have their systems fully under control, and so are generating data that do not represent competent performance, the outlier scheme may fail to identify all the outliers. Under such conditions, the robust estimates are clearly superior to the traditional estimates since the robust estimates are not influenced by persistent outliers. Because they are universally appropriate, the robust estimates of the mean and standard deviation are used to characterize acceptable performance.

The remainder of the statistical report is a historgram and an ordered listing of all the data from USEPA and state laboratories with outliers followed by an asterisk. The histogram provides a convenient visual impression of how the retained data are distributed and the ordered listing is necessary to document the actual data that the statistics were developed from. Figure 1 is an example of the report.

2.2 Developing Regressions from Previous Study Statistics

Where there are no available historical results for use in setting limits, either because the samples for previous studies were made differently or did not include particular chemical constituents, limits must be calculated directly from the specific USEPA and state laboratory statistics for that sample. If, however, appropriate background statistics are available, estimates of the mean and standard deviation for the current study samples can be made from linear regressions generated from the historical statistics.

These linear regressions, for $\bar{X}$ versus true value (T) and S versus T, are achieved by using the customary least-squares algorithm to fit:

$$\frac{\bar{X}}{T} = a + b\left(\frac{1}{T}\right) \quad \text{and} \quad \frac{S}{T} = c + d\left(\frac{1}{T}\right) \tag{5}$$

which can then be converted to the desired relationships by multiplying through by T, giving:

$$X = aT + b \quad \text{and} \quad S = cT + d. \tag{6}$$

This technique leads to a minimization of the sum of the squares of the residuals as a percentage of the true value associated with each residual. If the least-squares algorithm is used to fit the desired relationship directly, the statistics for the samples with high true values will tend to dominate the regressions. This can, in turn, cause regressions which provide ridiculous estimates of the mean and standard deviation for samples with low true values and, therefore, must be avoided. This regression procedure was selected after a comparison of results with traditional linear, quadratic, cubic and power curve fits to the historical statistics.

2.3 Establishing the Limits

Regression estimates of the mean and standard deviation should be compared with the statistics calculated from the study data. Since general comparability is all that is needed, a visual comparison is considered adequate. If both sets of statistical estimates agree reasonably well, the regression estimates should be used as the basis for acceptance limits because they are independent of the data to be evaluated, there is more information behind them and, for those studies involving two samples of each chemical constituent, the regression equations force a smooth and logical relationship between the statistical estimates for the two samples.

If the predicted statistics (mean and standard deviation) differ from the corresponding sample statistics, the regression estimates are usually inappropriate for some reason. Such a problem may result from the use of regression equations to estimate statistics at true values outside the concentration range from which the regression were developed, use of incorrect true values or a significant change in the way the samples were made. If suitable regression estimates cannot be developed, the study statistics themselves must be used as the basis for acceptance limits.

3. PERFORMANCE EVALUATION REPORTS

Once appropriate statistics are available, it is a simple matter to calculate the 95 or 99 percent confidence limits to be used as acceptance limits and proceed with evaluating the study data. An individual report is generated for each participating laboratory showing the data it reported, the related true values and limits, and an evaluation judgment for each reported value based upon its relationship to the limits. Each laboratory receives a copy of its report and a copy is sent to the responsible USEPA regional office or state for follow-up contact, as necessary, to resolve unacceptable performance.

4. DISCUSSION AND CONCLUSIONS

To date, USEPA has successfully completed 16 studies involving a total of 93 water contaminants and thousands of participating laboratories. The results of our studies have been well received by the participating laboratories, the USEPA regional offices and the states that depend on our reports to highlight the laboratories in their area requiring priority attention. Where states have a similar interest in evaluation of laboratory performance, it is increasingly common for them to request that their laboratories be included in our studies rather than developing a system of their own.

The current procedures for establishing the acceptance limits, particularly how to estimate study statistics and fit the regressions, have developed over time as shortcomings in each succeeding system have become apparent. There have been no significant changes for several years and none are anticipated in the future. The current limits seem to do a satisfactory job of characterizing the range of data from a competent water laboratory operating within state-of-the-art capabilities.

REFERENCES

- Grubbs, Frank E., 1950. Sample Criteria for Testing Outlying Observations, Annals of Mathematical Statistics, Vol. 21, pp. 27-55.

- Grubbs, Frank E., 1969. Procedures for Detecting Outlying Observations in Samples, Technometrics, Vol. 11, No. 1, Feb., pp. 1-21.

- American Society for Testing and Materials (ASTM), Subcommittee E-11. E178-Recommended Practice for Dealing with Outlying Observations, Annual Book of ASTM Standards, Part 41.

- Dixon, W. J. and F. J. Massey, Jr., 1969. Introduction to Statistical Analysis, Third Ed. McGraw-Hill Book Co., New York, NY, pp. 330-331.

- Stevens, M. A., 1974. EDF Statistics for Goodness of Fit and Some Comparisons, Journal of the American Statistical Association, Vol. 69, No. 347, Sept., pp. 730-737.

- Green, J. R. and Y. A. S. Hegazy, 1976. Powerful Modified EDF Goodness-of-Fit Tests, Journal of the American Statistical Association, Vol. 71, No. 353, March, pp. 204-209.

WP006 ALUMINUM SAMPLE 1 TRUE VALUE = 5040.

N	MEAN	STD. DEV.	LOWER LIMIT	UPPER LIMIT
57	4902.0701	1105.4277	1008.0000	25200.0000
55	5070.4000	664.8220	3295.1923	6845.6076
52	5204.6538	359.2784		

25% WINSORIZED MEAN = 5155.3077
STANDARD DEVIATION ESTIMATED FROM IQD = 296.5159

REGARDING NORMALITY OF THE MIDDLE 50% OF THE DATA:

THE KOLMOGOROV - SMIRNOV STATISTIC, D = .0718
D > .1768 INDICATES NON-NORMALITY AT THE 5% SIG. LEVEL.
D > .2044 INDICATES NON-NORMALITY AT THE 1% SIG. LEVEL.

REGARDING NORMALITY OF THE DATA AFTER THE TRADITIONAL REMOVAL OF OUTLIERS:

THE KOLMOGOROV - SMIRNOV STATISTIC, D = .1049
D > .1223 INDICATES NON-NORMALITY AT THE 5% SIG. LEVEL.
D > .1414 INDICATES NON-NORMALITY AT THE 1% SIG. LEVEL.

THE ANDERSON - DARLING STATISTIC, A = .8852
A > .7371 INDICATES NON-NORMALITY AT THE 5% SIG. LEVEL.
A > 1.0228 INDICATES NON-NORMALITY AT THE 1% SIG. LEVEL.

MINIMUM

VALUE	FREQ.	
4300.0000	2	00
4545.0000	1	0
4790.0000	16	XXXXXX0000000000
5035.0000	13	XXXXXXXXXXXX
5280.0000	12	XXXXXXX00000
5525.0000	5	00000
5770.0000	1	C
6015.0001	2	00

NON-BLANK LAB DATA VALUES:

SEQ.	VALUE	SEQ.	VALUE	SEQ.	VALUE	SEQ.	VALUE
47	1.0000*	23	4980.0000	94	5100.0000	63	5400.0000
98	545.0000*	8	5000.0000	38	5140.0000	89	5400.0000
20	2580.0000*	55	5000.0000	101	5147.0000	26	5400.0000
60	2650.0000*	41	5000.0000	7	5160.0000	43	5500.0000
99	3000.0000*	42	5000.0000	92	5170.0000	19	5500.0000
22	4300.0000	45	5000.0000	5	5200.0000	82	5500.0000
13	4500.0000	40	5000.0000	6	5200.0000	105	5600.0000
78	4700.0000	108	5001.0000	58	5220.0000	39	5620.0000
14	4810.0000	48	5030.0000	77	5300.0000	37	5665.0000
36	4870.0000	3	5050.0000	111	5300.0000	1	5740.0000
44	4900.0000	12	5090.0000	17	5300.0000	96	5750.0000
73	4900.0000	11	5100.0000	112	5300.0000	83	5900.0000
18	4900.0000	97	5100.0000	95	5310.0000	100	6039.0000
103	4910.0000	87	5100.0000	21	5320.0000	109	6260.0000
62	4960.0000						

Figure 1. Example of the Statistical Report for Each Constituent in a Sample